EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

Organised by

Universitat d'Alacant

Universidad de Alicante

transducens

research group

ISBN: 978-84-09-01901-4

Europarl Datasets with Demographic Speaker Information

Eva Vanmassenhove
ADAPT Centre
School of Computing and Engineering
Dublin City University
Dublin, Ireland
eva.vanmassenhove@adaptcentre.ie

Christian Hardmeier
Department of Linguistics and Philology
Uppsala universitet
Uppsala, Sweden
christian.hardmeier@lingfil.uu.se

1 Problem Statement

Research on speaker-adapted neural machine translation (NMT) is scarce. One of the main challenges for more personalized MT systems is finding large enough annotated parallel datasets with speaker information. Rabinovich et al. (2017) published an annotated parallel dataset for EN–FR and EN–DE, however, for many other language pairs no sufficiently large annotated datasets are available.

2 Datasets

To address the aforementioned problem, we publish a collection of parallel corpora licensed under the Creative Commons Attribution 4.0 International License for 20 language pairs available online: https://github.com/evavnmssnhv/Europarl-Speaker-Information. We tagged parallel sentences from Europarl (Koehn, 2005) with speaker information (name, gender, age, date of birth, euroID and date of the session) based on monolingual Europarl source files which contain speaker names on the paragraph level. We used meta-information of the members of the European Parliament (MEPs) released by Rabinovich et al. (2017) to retrieve the demographic annotations. An overview of the language pairs as well as the amount of annotated parallel sentences per language pair is given in Table 1.

3 Analysis

Additionally, we analyzed the EN–FR dataset with respect to the percentage of male versus female speakers in various age groups (see Figure 1).

Languages	# sents	Languages	# sents
EN–BG	306,380	**EN–IT**	1,297,635
EN–CS	491,848	**EN–LT**	481,570
EN–DA	1,421197	**EN–LV**	487,287
EN–DE	1,296,843	**EN–NL**	1,419,359
EN–EL	921,540	**EN–PL**	478,008
EN–ES	1,419,507	**EN–PT**	1,426,043
EN–ET	494,645	**EN–RO**	303,396
EN–FI	1,393,572	**EN–SK**	488,351
EN–FR	1,440,620	**EN–SL**	479,313
EN–HU	251,833	**EN–SV**	1,349,472

Table 1: Overview of annotated parallel sentences per language pair

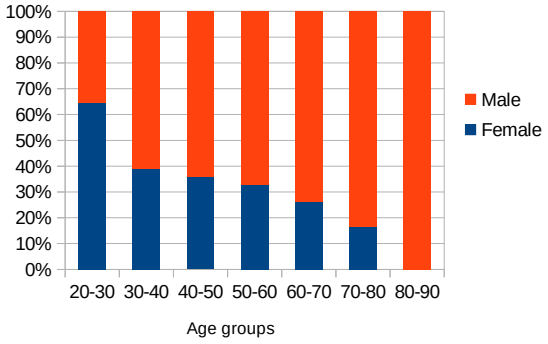

Figure 1: Percentage of female and male speakers per age group

References

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia and Shuly Wintner, 2017. Personalized Machine Translation: Preserving Original Author Traits. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain, 1074–1084.

Philipp Koehn, 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the In MT Summit*, Phuket, Thailand, 79–86.

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 371
Alacant, Spain, May 2018.